# マルチスケール・ブートストラップによる<br>近似的に不偏な selective inference

大阪大学 大学院基礎工学研究科，理化学研究所 革新知能統合研究センター 寺田 吉壱
京都大学 大学院情報学研究科，理化学研究所 革新知能統合研究センター 下平 英寿

データ解析において多くの統計的推測は，選択的な状況で行われている．例えば，階層的クラスタリングにおける各クラスタに関する検定では，予め仮説を用意するのではなく，データから得られたクラスタに対して検定を行うことが多い．近年，このように選択的な状況下において統計的推測を行う枠組みとして，selective inference が注目されている．本発表では，マルチスケール・ブートストラップを用いて，一般的な状況で近似的に不偏な selective inference を行う方法を提案する．

本稿では，選択的状況としてデータが帰無仮説を表現する領域の外側にある状況を想定した selective inference を考える．これは，あるクラスタが得られたときにのみ，そのクラスタの否定を帰無仮説とした検定を行う状況を想定している．ブートストラップ法では，データ $\mathcal{X} = (x_1, \ldots, x_n)$ から復元抽出により $\mathcal{X}^* = (x_1^*, \ldots, x_{n'}^*)$ を生成する．Shimodaira (2008) と同様に，データの何らかの変換 $f(\mathcal{X}) =: y \in \mathbb{R}^{m+1}$ を用いて，近似的に

$$Y \sim N(\mu, I),\ Y^* \sim N(y, \sigma^2 I),\ \sigma^2 = \frac{n}{n'}$$

とできることを仮定する．また，$y = (y_1, \ldots, y_{m+1}) \in \mathbb{R}^{m+1}$ に対して，$u = (y_1, \ldots, y_m)$，$v = y_{m+1}$ とおき，領域 $\mathcal{S} := \{(u, v) \mid v > -h(u),\ u \in \mathbb{R}^m\}$ に対して，$y \in \mathcal{S}$ が選択的状況を表すとする．このとき，$\mathcal{H} = \mathcal{S}^c$ を仮説領域として，$\mu \in \mathcal{H}$ を帰無仮説とする selective inference を考える．ここで，境界 $\partial\mathcal{H}$ を表す連続関数 $h(u)$ は，Shimodaira (2008) の意味で nearly flat であると仮定する．すなわち，$h$ とそのフーリエ変換 $\tilde{h}$ の $L^1$-ノルムが有界（$\|h\|_1, \|\tilde{h}\|_1 < \infty$）であると仮定し，$h$ の $L^\infty$-ノルム $\|h\|_\infty = O(\lambda^2)$ と表すときに，$\lambda \to 0$ のときの漸近理論を考える．

ブートストラップ確率 $P_{\sigma^2}(Y^* \in \mathcal{H} \mid y)$ を $\alpha_{\sigma^2}(\mathcal{H} \mid y)$ と記し，$z_{\sigma^2}(\mathcal{H} \mid y) := \Phi^{-1}(1 - \alpha_{\sigma^2}(\mathcal{H} \mid y))$ を定義する．そして，$\sigma z_{\sigma^2}(\mathcal{H} \mid y)$ が $\beta$ をパラメータとするモデル $\psi_{\mathcal{H}}(\sigma^2 \mid \beta)$ によって（近似的に）表現されるとする．実際に，$\partial\mathcal{H}$ が滑らかであれば，$\sigma z_{\sigma^2}(\mathcal{H} \mid y) = \beta_0 + \beta_1\sigma^2 + \beta_2\sigma^4 + \cdots$ と表すことが出来る．このとき，selective inference の $p$ 値として，

$$p(\mathcal{H} \mid y) := \frac{1 - \Phi(\psi_{\mathcal{H}}(-1 \mid \beta(y)))}{\Phi(\psi_{\mathcal{H}}(0 \mid \gamma(y)) - \psi_{\mathcal{H}}(-1 \mid \beta(y)))}$$

を提案する．ここで，$\beta(y)$ は $\psi_{\mathcal{H}}(\sigma^2 \mid \beta(y)) = \sigma z_{\sigma^2}(\mathcal{H} \mid y)$ を満たすパラメータである．適当な正則条件の下で，

$$\forall y \in \partial\mathcal{H};\ \frac{P(p(\mathcal{H} \mid Y) < \alpha \mid y)}{P(Y \in \mathcal{H}^c \mid y)} = \alpha + O(\lambda^2)$$

が成り立つので，$p(\mathcal{H} \mid y)$ を用いた検定は近似的に不偏な selective inference となる．実際には，パラメータ $\beta(y)$ の値は未知であるから，いくつかの $n'$ に対してブートストラップを行うことで $\beta(y)$ を推定し，$\beta(y)$ をその推定量 $\hat{\beta}(y)$ に置き換えて $p(\mathcal{H} \mid y)$ を計算することで，近似的に不偏な selective inference を実行することができる．

一方で，$\partial\mathcal{H}$ がなめらかでない場合を想定したモデル $\psi_{\mathcal{H}}(\sigma^2 \mid \beta)$（例えば，$\psi_{\mathcal{H}}(\sigma^2 \mid \beta) = \beta_0 + \beta_1\sigma$）では，$\sigma^2 = -1$ へ直接的に外挿することはできない．このような場合，Shimodaira (2008) と同様に，$\sigma^2 = \sigma_0^2 > 0$ における $\psi_{\mathcal{H}}(\sigma^2 \mid \beta)$ のテイラー展開を $k$ 項で打ち切った関数を $\psi_{\mathcal{H}}$ の代わりに用いて $\sigma^2 = -1$ へ外挿する．このアプローチの正当性を含む理論の詳細や提案手法を階層的クラスタリングの信頼性評価に応用した例については当日報告する．

## 参考文献

[1] Shimodaira, H. (2008). Testing Regions with Nonsmooth Boundaries via Multiscale Bootstrap. *Journal of Statistical Planning and Inference*, **138**, 1227–1241.